

II

ALIGNMENTS


Part II of this compendium contains alignments of PV coding sequences and their corresponding protein sequences. You can find your way in this section by looking at either the headline or the page number, both of which contain the gene name. Unless there is evidence for believing that a coding sequence begins later in the open reading frame, coding sequences are presented from the first methionine codon of the reading frame, and protein sequences from the corresponding methionine.

The order of the alignments is E6, E7, E1, E2, E4, E5, L2, L1, L1 Consensus Primer Region, and LCR. The L1 Consensus Primer Region alignment includes an expanded set of PV types, as well as many novel sequences that will probably attain type status in the near future. Many of these sequences have been sequenced only over this fragment of L1. The LCR was operationally defined as the region after the end of L1 and before the first methionine of the E6 ORF. The LCR region is presented as a nucleotide alignment only, and contains only HPVs (no animal PVs).

This year, the sequences have been grouped within the alignments according to the taxonomic system proposed in Chan et al., *J. Virol.* **69**:3074–83. Under this system, sets of relatively close sequences are termed “groups”, while sets of relatively close “groups” are termed “super-groups”. Each of the groups has at least two members, although in some cases, e.g. GroupA3 and GroupA11, only one member of the group has been sequenced over its complete genome, so that the group will have only one representative in many of the alignments. Some sequences clearly belong to a particular supergroup but are not related closely enough to any other sequence to justify including them in a group. In addition, some few PV sequences (FPV, MmPV, MnPV) are so distant from all other PVs that they cannot be considered as belonging to any of the supergroups. These sequences have been placed together at the very bottom of the alignments.

At the head of each group is a consensus sequence for that group. Within the consensus, capitalized letters indicate that the base or amino acid is present at that location in all taxa in the group, that is to say, it is completely conserved; lower case letters indicate that the base or amino acid is present in 50% or more of the group sequences; a question mark indicates that no one base or amino acid is present at that position 50% or more of the time. Each supergroup consensus sequence represents a consensus for all sequences in that supergroup (not including the individual group consensus sequences). Each supergroup consensus is placed at the head of the groups which are its members. Immediately following the supergroup consensus are sequences belonging to the supergroup, but not assigned to any group, e.g. HPV54. A consensus sequence for those sequences belonging to no supergroup is given as “Unclass.con”

Each set of sequences is referenced to the consensus sequence immediately above them. Agreement with the consensus sequence at any location is shown by a dash (-) while gaps are indicated by dots (. . .). Blank spaces within the alignment indicate lack of sequence information over that region. Occasionally, a nucleotide sequence will contain a percentage sign, (%), which indicates that the sequence appears to contain a frameshift indel at that position.

Typically, the alignments are displayed so that all the sequences on the same page, as well as the sequences on the facing page, are homologous to one another. There are two exceptions to this general rule. At the beginnings and ends of some of the alignments, some sequences may be separated by lines of arrows (➡). This indicates that these sequences are significantly longer than the rest, and that they continue below on the same page. The other exception is found in the E4, E5 and LCR alignments. These differ from the other alignments in two respects: first, sequences on facing pages are not homologous to each other; second, these regions contain solid “separation bars” () throughout the alignment, to indicate that no significant similarity exists between sequences above and below the bar, and they could not be aligned for that reason. In some cases, it seems probable that this lack of similarity may be attributable to an absence of homology between the sequences.

The alignments have been created progressively, beginning with pairwise alignments within the groups and then proceeding to inter-group comparisons.

Contents

PART II Alignments

Introduction	II-1
Contents	II-2
E6 Protein Alignment	II-E6-2
E6 Nucleotide Alignment	II-E6-9
E7 Protein Alignment	II-E7-2
E7 Nucleotide Alignment	II-E7-8
E1 Protein Alignment	II-E1-2
E1 Nucleotide Alignment	II-E1-22
E2 Protein Alignment	II-E2-2
E2 Nucleotide Alignment	II-E2-18
E4 Protein Alignment	II-E4-2
E4 Nucleotide Alignment	II-E4-10
E5 Protein Alignment	II-E5-2
E5 Nucleotide Alignment	II-E5-4
L2 Protein Alignment	II-L2-2
L2 Nucleotide Alignment	II-L2-22
L1 Protein Alignment	II-L1-2
L1 Nucleotide Alignment	II-L1-21
L1 Consensus Primer Region Protein Alignment	II-L1 CPR-2
L1 Consensus Primer Region Nucleotide Alignment	II-L1 CPR-8
LCR Nucleotide Alignment	II-LCR-2